# Ex Post Stability of Constitutions: an Endogenous Explanation

*Lorenzo Sacconi*[1] and *Virginia Cecchini Manara*[2]

*1 – Department of Economics and Management, University of Trento and EconomEtica, Italy*

*(e-mail: lorenzo.sacconi@unitn.it)*

*2 – Institute of Economics, Sant'Anna School of Advanced Studies, Pisa, Italy*

*(e-mail: v.cecchinimanara@sssup.it)*

---

***Abstract***

*In the tradition of Constitutional Political Law, the State arises as an institutional means for resolving interpersonal conflicts. In Buchanan's view, this institution entails a two-stage contractual process: a constitutional contract that defines a structure of rights via mutual agreement, and a second stage where, once individual rights are acknowledged, contractual negotiations become possible (post constitutional contracts). In order for this institution to be viable, one should also recognize the necessity of some enforcing agent for the protection of individual rights to do things, including the making and carrying out of valid contracts. Thus the Constitution should ex ante commit to punish any ex post transgression of post constitutional agreements, but then we face the problem of the abuse of authority: no formal commitment can prevent the State from abusing, and the presence of constitutional limits is not enough.*

*If we model the pre-agreement state of nature as a Prisoners' Dilemma or a Free-Rider Problem, then each individual has a rational incentive to defect, which leads to a suboptimal outcome; this situation justifies the mutual agreement of the social contract for a cooperative interaction, but at the same time and for the same reason this agreement is not stable, since non-compliance (naturally corresponding to mutual defection) is the only equilibrium.*

*How does it happen that we observe citizens acting against the abuse of authority by those who hold political power, even if this behavior is not in their own material interest?*

*In this paper we provide a game theoretical model of the interplay between the State and the citizens in terms of a repeated Trust Game with psychological Nash equilibria (Geanakoplos, Pearce and Stacchetti, 1989; Rabin, 1993) and according to the theory of conformist preferences (Grimalda and Sacconi, 2005).*

*We argue that a Rawlsian Social Contract (Rawls, 1971) is able not only to solve the normative equilibrium selection problem, i.e. to choose a constitutional order through a decision procedure that satisfies elementary conditions of impersonality, impartiality, and empathy, as Binmore (2005) has shown, but it can also solve the problem of the ex post stability of the Constitution through the formation of endogenous motivations, suggesting an illuminating explanation of why (sometimes) some of us comply with just institutions even if we have some direct material incentive not to do so.*

## 1. Introduction

In the tradition of Constitutional Political Law, the State arises as an institutional means for resolving interpersonal conflicts. In fact, when separate persons and groups have conflicting claims, most of the problems of social interaction arise and the situation would soon fall into a Hobbesian anarchy, unless we find some institutionalized means of resolving interpersonal disputes[1].

In James Buchanan's view, this institution entails a two-stage contractual process: a *constitutional contract* that defines a structure of rights via mutual agreement, and a second stage where, once individual rights are acknowledged, contractual negotiations become possible (*post constitutional contracts*). In order for this institution to be viable, one should also recognize the necessity of some enforcing agent, a collectivity, a state, for the protection of individual rights to do things, including the making and carrying out of valid contracts. Thus the Constitution should *ex ante* commit to punish any *ex post* transgression of post constitutional agreements.

We can find here a fundamental problem of *stability* of the Constitutional Contract: if the state of nature that precedes the agreement on a Constitution is analogous to the classical Prisoners' Dilemma in game theory, then, as Buchanan himself admitted, "any positive structure of rights is extremely vulnerable to defection if continued adherence to the contractual basis depends on voluntary and independent «law-abiding»"[2].

This problem can be better understood in the light of the distinction proposed by David Gauthier[3] between the *ex ante* and the *ex post* perspective of an impartial agreement (i.e. the social contract). The *ex ante* decision problem (why to enter the social contract) poses a question of **justification** of the agreement, while the *ex post* decision problem (why to comply with the social contract) concerns its stability and **compliance**.

If we model the pre-agreement state of nature as a Prisoners' Dilemma or a Free-Rider Problem, then each individual has a rational incentive to defect, which leads to a suboptimal outcome; this situation *justifies* the mutual agreement of the social contract for a cooperative interaction, but at the same time and for the same reason this agreement is not *stable*, since non-compliance (naturally corresponding to mutual defection) is the only equilibrium.

[1] See Buchanan (1975), chapter 1.
[2] Buchanan (1975), 7.4.29.
[3] See Gauthier (1986), pp. 116-8.

The idea of Buchanan is that, in order to make the agreement stable *ex post*, it is enough to delegate the authority to enforce the contract to an external agency: "If individual parties to an initial contract in which property assignments are established mutually acknowledge the presence of incentives for each participant to default and, hence, recognize the absence of viability in any scheme that requires dependence on voluntary compliance, they will, at the time of contract, enter into some sort of enforcement arrangement. Individuals' claims to stocks of goods and endowments will be accompanied by some enforcement institution that will be aimed to secure such claims. The nature of this enforcement contract or institution must be carefully examined. Each person will receive some benefit from the assurance that his established claims will be honored by others in the community. And there are mutual gains to all parties from engaging in some joint or collectivized enforcement effort. Enforcement of property claims, of individuals' rights to carry out designated activities, qualifies as a «public good» in the modern sense of this term. (…) This problem may be handled by an agreement by all persons on the purchase of the services of some external enforcing agent or institution that will, in all particular cases, take the enforcing-punishment action required"[4].

This solution poses the problem of the *Abuse of Authority*, which was clear also in Buchanan's work, and can be summarized under the label *Quis Custodiet Ipsos Custodes?*: "if the State is empowered to enforce individual rights, how is it to be prevented from going beyond these limits? What are the «rights» of the enforcing agent itself, the state? … How can Leviathan be chained?"[5]. The idea that the limitations to the power of the State can be written in the Constitution and solved by a constitutional *ex ante* commitment is in contradiction with what we know from non-cooperative game theory, where agreements are not enforceable and commitments are not binding. If the Abuse of Authority is incentive compatible for the State, no formal commitment will prevent him from abusing, and the presence of **constitutional limits** is not enough[6].

In this paper we propose that a Rawlsian Social Contract (Rawls, 1971) is able not only to solve the *normative* equilibrium selection problem, i.e. to choose a constitutional order through a decision procedure that satisfies elementary conditions of impersonality, impartiality, and empathy, as Binmore (2005) has shown, reevaluating John Rawls' egalitarian and maximin principle of justice within a game theoretical perspective. But it can also solve the problem of the *ex post* stability of the Constitution and of the Abuse of Authority through the formation of endogenous

[4] Buchanan (1975), 7.4.34-36.

[5] Buchanan (1975), 7.1.31.

[6] For the discussion of this "contractarian compliance impossibility", see Sacconi, Faillo and Ottone (2011).

motivations, not only through formal constitutional limits, suggesting an illuminating explanation of why (sometimes) some of us comply with just institutions even if we have some direct material incentive not to do so.

Justice as fairness, Rawls says, understood as the set of principles of justice chosen "under a veil of ignorance" – once the principles are assumed to shape the institutions of a well-ordered society – provides its own support to the stability of just institutions. In fact when institutions are just (here it is clear that we are taking the *ex post* perspective, i.e. once the constitutional decision from the *ex ante* position has already been taken and for some reason has been successful), those who take part in the arrangement develop a sense of justice that carries with it the desire to support and maintain that arrangement. The idea is that motives to act are now enriched with a new motivation able to overcome the counteracting tendency to injustice. Note that instability is clearly seen in term of a Prisoner's Dilemma-like situation: institutions may be unstable because complying with them may not result in the best response of each participant to other members' behavior. However, the sense of justice, once developed, overcomes incentives to cheat and transforms fair behavior into each participant's best response to the other individuals' behaviors.

The relevant features of Rawls' theory are captured in the conformist preferences model (Grimalda and Sacconi 2005; Sacconi and Grimalda 2007) which is based on a different notion of equilibrium – the psychological Nash equilibrium (see Geanakoplos, Pearce and Stacchetti, 1989; Rabin, 1993).

Binmore (2005) has shown that in a context of constitutional choice where agents confront one another in a state of nature, an egalitarian constitution is agreed. By *constitutional choice* is meant the selection of an admissible subset of the players' state of nature strategies, that if unconstrained would allow them to undertake any opportunistic behavior in their relationships. Under the ethical assumption of the veil of ignorance, however, they reach agreement on a constitution such that they make a final allocation of payoffs which is identical to the best *egalitarian* distribution of the cooperative surplus among those feasible for the constitutional choice.

In order for such equilibrium selection device to be effective in solving the compliance problem related to the social contract, however, it is required that also *ex post*, when players exit the veil of ignorance, or the original position wherein they select the egalitarian solution under the veil, they hold shared (if not common) knowledge that all of them will abide by their strategy components of the agreed equilibrium. However, when players leave the original position under the

veil of ignorance and resume their real-life interaction, it is no longer clear that the equilibrium selection under the veil is still convincing enough to induce their spontaneous adherence *ex post*, beyond the veil. Of course, the agreed solution is still a (feasible) equilibrium point *ex post*, but this may not be sufficient to assure actual coordination on the egalitarian equilibrium (which does not entail that there are no cases in which this may work). In order to assure it, what is required is the actual knowledge of other players behavior, but this cannot be derived – at least cannot be logically deduced – from the mere hypothesis that they would have agreed unanimously on the selection of an equilibrium point under the veil of ignorance

In fact, when the veil is lifted and they return to the game of life, they must consider the basic state-of-nature game, and if the state of nature is a repeated game, they have gone back to the initially possible set of equilibria. In the game of life, players may doubt that others will continue to select the egalitarian solution. They face the old repeated game of life, not anymore the bargaining game under condition of impartiality (and the symmetrized set of payoffs). In this game they cannot say of knowing (in the epistemic sense of knowing the truth) that they all will play the *ex ante* selected equilibria.

Notice that, in order to explain the role of the social contract, we take the game theoretic perspective of a repeated game between those who occupy positions of authority and citizens who are subject to this authority: the interplay between the State and citizens is described in terms of a repeated Trust Game.

Although an *ex ante* social contract would be able to justify the choice of a fair equilibrium, *ex post* we are faced with the problem of the incentives to which players will respond when they exit from the original-position-and-veil-of-ignorance thought experiment and return to “the game of life”, where they play according to the entire set of their preferences and motivations to act. The problem is that in the presence of multiple equilibria, each with some motivating force conditional on existence of a system of expectations consistent with it, no particular equilibria has definitive reason to be carried out, and thus the one corresponding to the *ex ante* agreement need not have any incentive effect on compliance.

In fact the reputation equilibrium chosen by the State would not be the fully fair and cooperative one, but rather the one whereby the politicians acquire a reputation for abusing the trust of citizens – but only to the extent that makes them indifferent between maintaining their relations of cooperation and withdrawing from them (a “sophisticated” abuse).

There is some evidence of this behavior in real life relationships between citizens and those who are in the position of power. Politicians may comply in only few cases, or to a minimal extent, with their duties.

However, there is also evidence of citizens' activism that refuses to acquiesce and actively countervails such hypocritical political conduct. Many examples illustrate behaviors by active citizens that cannot be captured in terms of their mere self-interest and cannot be understood as mere defense of their own material interest.

Admittedly some of these behaviors can be understood as reflecting a concern for other citizens' well-being, rather than the well-being of the active citizens themselves. More exactly, however, they express the citizens' attachment to impersonal principles of justice, i.e. a desire to conform with socially accepted norms of fair treatment – even when such conformity concerns not so much the active citizen itself but mostly the well-being of third parties. Hence only disinterested (from the egoistic point of view) motivations may be of relevance in explaining such action.

Yet the egalitarian equilibrium does not necessarily become unstable. If players have rationally agreed on a certain constitution, this affects their preferences, admitted that they expect reciprocity from others in complying with the same institution. Under expectations of reciprocity, preferences incorporate a desire to conform with agreed principles, which is a direct function of the extent to which an agent is conforming given his expectation of others' behavior, and the extent to which he expects reciprocal conformity by others given their belief in his own action. Impartial agreement on a principle lays the bases; then mutual expectations are essential in shaping a preference for conformity. Moreover the mental experiment of an agreement under an impartial perspective may affect their beliefs, if not giving them exact knowledge. So, as is observed in experiments mimicking the social contract in the lab and related to conformist preferences (Sacconi, Faillo and Ottone, 2011), players that are back to the game of life beyond the hypothetical agreement may form by default their beliefs over any other player behavior consistently with the mental model of a player who agreed the social contract and hence at least expressed the intention or planned to act according to the social contract. As a fact of cognitive psychology, 'normally' players who plan an action will carry it out. This psychologically explains why they entertain first and second order beliefs of players' high level of consistency between actions and principles provided by the *ex ante* selected social contract. As far as these beliefs are realistic, the theory of conformist preferences causally provides a motivational explanation of why *ex post* players may

comply with the social contract even if they don't hold common knowledge of the ongoing game solution.

A proper understanding of these third-parties-concerned non-egoistic behaviors in terms of norm compliance based on conformist preferences has been the focus of previous works on this topic (see Grimalda and Sacconi, 2005; Sacconi and Grimalda, 2007, Sacconi and Faillo, 2010). Here we shall try to make sense of the evidence by focusing on the basic State/citizens bilateral strategic relationships.

The main result of this paper is that a social contract perspective helps reducing to *just two* the candidate reputation equilibria that *ex post*, in the real world interaction taking place beyond the "veil of ignorance", may be played. These equilibria coincide with *i)* complete compliance with the social contract and *ii)* the citizens punishing of the governors' incomplete compliance by staying out from the relation even when their material payoff would push them to give in to the State sophisticated abuse. Sophisticated abuse repeated equilibria do not belong anymore to the equilibrium set.

## 2. The Trust Game

To gain better understanding of where we stand, consider that the appropriate game representation of the State/citizens interaction is the iterated **Trust Game**, with the stage game illustrated in Figure 1.

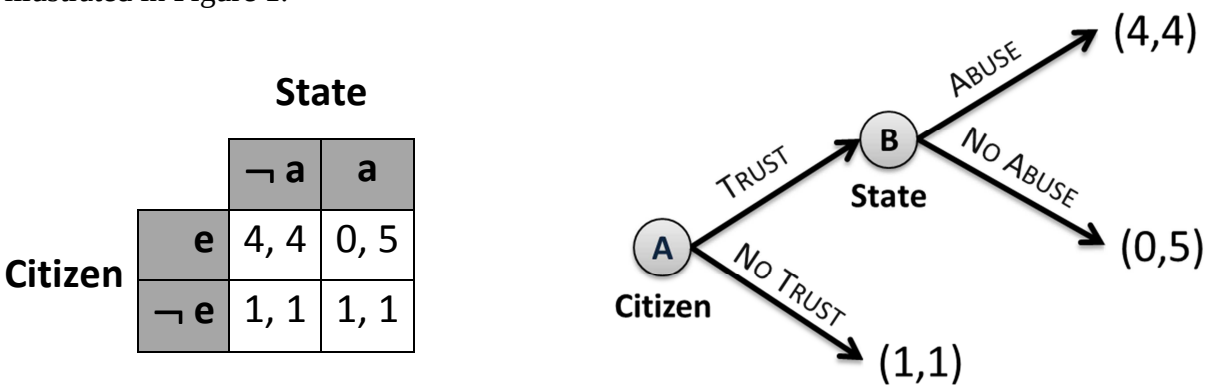


**Figure 1** *One-shot Trust Game in normal and extensive form*

Player A (the citizen) will enter (or not) by trusting (or not) player B (the State) and by participating to social cooperation with his own investment. Player B decides whether to appropriate player A's trust investment by abusing or not. If he chooses non-abuse the surplus from cooperation is shared in an equitable way. Otherwise the citizen is deprived of any benefit from entrance (including the endowment that s/he would possess if s/he did not invest in trust

relationships), while the part with authority gains a large profit. Note, however, that this mode of interaction is intuitively understood as socially inefficient in a utilitarian sense – that is, admitted utility comparability, the State still prefers individually to abuse, but the fair sharing in the case of non-abuse would yield a larger amount of interpersonal social welfare. However, notwithstanding any consideration of social efficiency, the only Nash equilibrium is the strategy pair such that B abuses and A stays out. The mutually beneficial outcome (4, 4) cannot be sustained in equilibrium as long as the game is played one shot.

But now consider the equilibrium set of the **repeated Trust Game** between the long-run State B, who receives the average payoff from all his participations into the infinite series of stage games, and the "average" citizen (call him/her again A because this is useful for considering the average payoff of an infinite series of short-run citizens that enter or otherwise the position of the one-shot A player at each repetition), who enters each stage game (or refuses to enter). Under the usual assumptions for reputation games, the repeated trust game will display a convex payoff space (constituted by all the average discounted payoff vectors obtainable from pairs of repeated strategies) coinciding with the convex envelope of the one-stage pure payoff vectors.

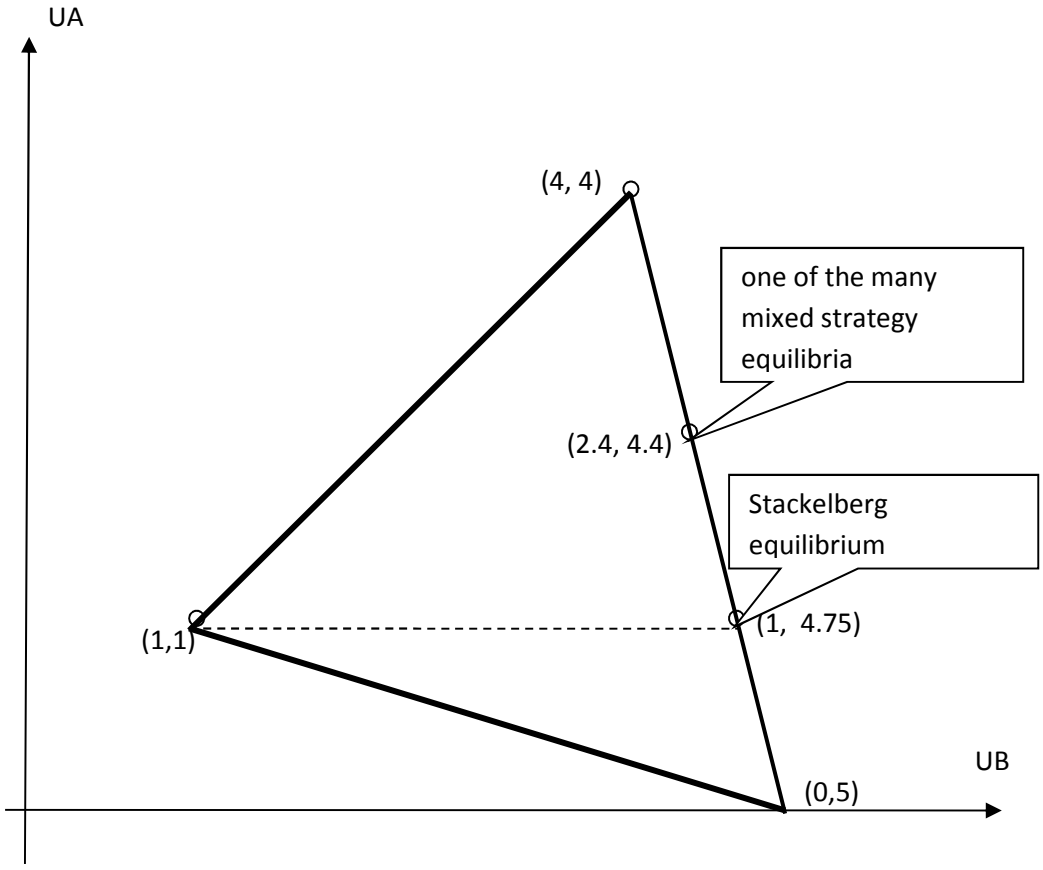

**Figure 2** *Repeated Trust Game*

Within this payoff space, every point above the dotted line corresponds to an equilibrium strategy profile such that player A 'enters' with a given frequency and player B abuses or not with the appropriate probability mixture (Fudenberg and Levine, 1989; Fudenberg 1991). Of course, the most relevant equilibria are those where player A never enters because player B will always abuse, with average discounted payoffs (1,1), and the equilibrium with average discounted payoffs (4,4) where player 2 never abuses and hence player 1 enters each time.

But it is also remarkable the *Stackelberg equilibrium*, where the State B is believed to make a commitment on the mixed strategy *(0.75**a**, 0.25 ¬**a**)*. In fact, B may develop a reputation for being this type by playing the two pure strategies with the attached probability throughout all the repetitions of the game. Thus each citizen in the role of player A necessarily enters, since his/her payoff is the same as staying out (namely 1) – i.e. he is indifferent between entering and staying out (if player B were to give him/her an infinitesimal additional positive utility ε by reducing his abuse probability correspondingly, 'entrance' would be certain). This gives B an average expected payoff of 4.75, which is the best payoff that player B can obtain in equilibrium. Then player B's best response is to stick to this type/commitment whenever he is able to convince player A that he is this type so that A responds with his/her best response to this type's mixed strategy.

Any equilibrium point exerts a (limited) motivational force able to command actual behavior, which is effective in so far as each player believes that other players will play their strategy components of the same equilibrium. One may wonder whether the fact that a norm has been agreed from an *ex ante* (pre-play) perspective and exhibits various levels of consistency with different equilibria, may affect the motivational force exerted by different equilibria in a game. A positive answer would amount to a restriction on the number of equilibrium points that have motivational force over the players' behavior. In other words, one may ask whether norms can 'refine' the equilibrium set of a game in terms of the motivational strength of certain equilibria over other equilibria.

We will now show how conformist preferences derived from the Rawlsian idea of a sense of justice may affect compliance with the constitutional social contract: preferences incorporating the sense of justice will affect compliance by selecting as admissible the only subset of equilibria which are compatible with compliance with the agreed principles.

This results from a modification of the players' utility functions through integration of preferences with an intrinsic component for norm compliance, seen not as unilateral and

unconditioned, but as conditioned by beliefs about other players' reciprocal conformity. The 'refinement effect' on the admissible equilibria that this change in the equilibrium notion entails is surprising (and unexpected). As we will see, the equilibrium set of the repeated Trust Games under this revision of the utility function shrinks dramatically to the pure strategy equilibria of the repeated psychological Trust Game.

## 3. The Rawlsian Sense of Justice

An original approach to the institutional compliance problem was suggested by John Rawls in the *Theory of Justice* (1971), where he proposed the "sense of justice" as a solution for the stability problem of a well-ordered society – i.e. a society whose institutions are arranged according to the principles of justice (norms in our sense) chosen under a "veil of ignorance".

To understand how this is possible, it is necessary to consider the definition of "sense of justice". Although it presupposes the development of lower-level moral sentiments of love and trust, understood as feelings of attachment to lower-level institutions (families and just associations), if these institutions are perceived to be just, it is noticeable that the sense of justice is a desire to act upon general and abstract principles of justice as such, once they have been chosen under a veil of ignorance as the shaping principles of institutions, and hence have proved beneficial to ourselves in practice. Note that it is not the case that we act upon the principles insofar as they are beneficial only to concrete persons with whom we have direct links and emotional involvements. Once the level of a morality of principles has been reached, our desire to act upon the principles does not depend on other people's approbation or on other contingent facts such as satisfaction of the interests of some particular concrete person. On the contrary, it is the system of principles of justice in itself that constitutes the object of the sense of justice.

The question to be answered thus becomes how it is possible that principles themselves are capable of influencing our affections - that is, of generating the sense of justice as a relatively self-contained "desire to conform with the principles". The answer is twofold.

First, the sense of justice is not independent of the *content* of principles. These are principles that we could have decided to agree upon under a veil of ignorance as expressions of our rationality as free and equal moral persons. These principles are mutually advantageous and hence impartially acceptable by a rational choice, even if it is made from an impartial perspective, for they promote our interests and hence have some relation with our affections (preferences). Thus, in order for a

sense of justice to develop, principles cannot be arbitrary. They must be those principles that would have been chosen by a rational impartial agreement.

Second, despite the intellectual effect of recognizing that principles are rationally acceptable, the basic fact about the sense of justice is that it is by nature a moral sentiment inherently connected to natural attitudes. Moral sentiments are systems of dispositions interlocked with the human capability to realize natural attitudes. Thus moral liability for lacking moral sentiments has a direct counterpart in the lack of certain natural attitudes which results in affective responses like a sense of guilt, indignation or shame. Hence, even though the thought experiment of a decision under the veil of ignorance merely aids us in the *intellectual* recognition of principles acceptability, the sense of justice retains a motivational force on its own, which can be only traced back to its nature as a moral sentiment or desire not entirely reducible to the experience of its intellectual justification.

The proper functioning of the sense of justice can be understood, however, as the third level of a process of moral learning which in its first two steps already cultivates moral sentiments of love for parents and trust and friendship vis-à-vis the members of just associations in which the individual already takes part – and which s/he re-elaborates on those pre-existing sentiments. "*Given that a person's capacity for fellow feeling has been realized by forming attachment in accordance with the first two …[levels] and given that a society's institutions are just and are publicly known to be just, then this person acquires the correspondent sense of justice as he recognized that he and those for whom he cares are the beneficiaries of these arrangements*" (Rawls, 1971, p. 491).

As seems clear, reciprocity is a basic element in this definition. In fact reciprocity is understood as a deep-lying psychological fact of human nature amounting to the tendency to "answer in kind". The sense of justice "*arises from the manifest intention of other persons to act for our good. Because they recognize they wish us well we care for their well being in return. Thus we acquire attachment to persons and institutions according to how we perceive our good to be affected by them. The basic idea is one of reciprocity, a tendency to answer in kind"* (Rawls, 1971, p. 494). Two aspects are to be noted concerning the other person's "manifest intention" which elicits the tendency to "answer in kind". We recognize the caring for our good deriving from other people acting consistently with the principles of justice. Hence reciprocity is elicited not from the mere coherence of institutions with the principles of justice, but from the fact that other people make our good by acting intentionally upon those principles. What matters is not just reciprocity in

accepting the principles, but the intention displayed by other players' concretely acting upon the principles for our well-being. Secondly, this intention cannot be a direct intention from concrete person toward us as particular persons. By complying with principles, our good is pursued in an unconditional way - that is, impersonally and not conditionally on any particular description of us based on contingent characteristics or positions.

## 4. The Model of Conformist Preferences

The key elements of Rawls's analysis have been incorporated into a formal model of **conformist preferences** (Grimalda and Sacconi 2005; Sacconi and Grimalda 2007), grounded on the literature on psychological games and reciprocity (Geanakoplos, Pearce and Stacchetti, 1989; Rabin, 1993). The main features of the model are summarized in the Appendix.

### a. Conformist preferences in the one-shot Trust Game

To begin, let us illustrate the conformist preference model with reference to its application to the one shot Trust Game involving the State and the citizen.

To calculate conformist psychological payoffs and equilibria, let's consider the game matrix of the stage game, reported in Figure 3 (that replicates Figure 1 for the reader convenience). There are four possible states of affairs $\sigma$ coinciding with the cells of the normal form matrix: $(\neg e, \neg a), (\neg e, a)$ with material payoffs (1,1); $(e, a)$ with material payoffs (0,5); and $(e, \neg a)$, with material payoffs (4,4).

When these states of affairs are qualified in terms of their consistency with an *ex ante* agreed ethical norm preference over them are *conformist* – where "consistency" is defined as how far the players' strategy choices (jointly a state) are from the set of actions that would completely fulfil the agreed ethical norm of equity. By norm we mean a principle of justice for the distribution of material utilities coinciding with the *ex ante* social contract.

Let us assume that players have agreed on a social contract concerning the principle of justice that should govern cooperation in society and that it prescribes to apply the Nash Bargaining Solution, which requires maximizing the product of individual surpluses net of the *status quo*.

In this particular case, the *status quo* coincides with the outcome of the no-entry strategy – (1,1) – which is the assurance level that player *A* can grant herself for whatever player B's choice, included the case that he doesn't start any trust based interaction. This pay-off must then be subtracted from whatever pay-off is used in the calculation of the Nash product annexed to any state

of affair (strategy combination). The two further matrices (see below) show respectively: the Nash bargaining product calculated for each pure strategy combination needed to measure the consistency of each state with respect to the principle *T* and the players' relevant degrees of conditional and expected reciprocal conformity for each state (Figure 4), and the overall pay-offs resulting from the addition of the psychological conformist preference weight $\lambda = 2$ to the material pay-offs where this addition is appropriate (Figure 5).

| | ¬ a | a |
|---|---|---|
| **e** | 4, 4 | 0, 5 |
| **¬ e** | 1, 1 | 1, 1 |

**Figure 3** *Trust Game in normal form*

| | ¬ a | a |
|---|---|---|
| **e** | (4-1)(4-1) = 9 | 0-1)(5-1) = - 4 |
| **¬ e** | (1-1)(1-1) = 0 | (1-1)(1-1) = 0 |

**Figure 4** *T values at each state*

| | ¬ a | a |
|---|---|---|
| **e** | (4+λ) = 6, (4+λ) = 6 | 0, 5 |
| **¬ e** | 1, 1 | (1+λ) = 3, (1+λ) = 3 |

**Figure 5** *Psychological Trust Game with conformist utilities included with $\lambda = 2$*

In order to understand the psychological payoffs reported in Figure 5, consider that if a player cannot do anything better to improve the "collective" value of the principle *T* with respect to the *status quo* by means of her/his unilateral decision given the expected strategy choice of the other player, then s/he will be considered completely compliant by choosing to keep the *status quo* (no deviation from maximal conformity can be ascribed to her/his responsibility since her/his choice cannot do any better to maximize *T* than keeping to the status quo). This feature of the model depends on considering compliance in a non-cooperative *ex post* context wherein players are able to deviate unilaterally from an agreed norm, and secondly by considering conformity as conditional on the other player's expected level of compliance. Hence, in cases like the Trust Game, if the State is expected to abuse, the citizen cannot do anything to improve the value of *T* on the status quo and therefore he will be considered fully compliant with the principle by deciding to stay out (as a matter of fact she could only worsen the *T* value by entering). At the same time, the State predicting that the citizen will stay out – given he believes that the State shall abuse – cannot modify the value of *T* with respect to the *status quo.* Thus whatever the State's strategy choice, it is fully compliant in this case. The result is that also in the (*no-entry, abuse*) equilibrium point of the basic Trust Game, the conformity weight $\lambda$ adds to the players' pay-offs. Under this respect, there is no difference between the case (*no-entry, abuse*) and the case of the citizen entering because he predicts that the State is going not to be abusive and the State refraining from being abusive because it predicts that the citizen will enter (*entry, no-abuse*) – which is obviously the case in which both players unconditionally maximize *T* and hence necessarily the weight $\lambda$ enters their payoffs as they are full compliant.

By contrast, if the citizen enters when the State is unilaterally predicted to abuse, she would minimize *T* with reference to the alternative choice open to her of not entering, which scores a higher level of *T*. At the same time, the State misses the opportunity to maximize *T* given the citizen's decision to enter, and hence the latter will be considered as not complying at all. This

implies that when the State unilaterally and successfully abuses its citizen, none of the conformist preferences can add value to the players' material pay-offs.

Lastly, if the State chooses a mixed strategy whereby the citizen's decision between entry or non-entry has no influence on the *T* value, the citizen, whether she decides to enter or not, would be unable to improve the value of *T*. Therefore, by staying out she maximizes *T* as well. If, however, the citizen still stays out, no State's strategy can do any better in maximizing *T* than the one just described, and thus the State is as well completely compliant as when it abuses. Hence, a State's mixed strategy responded to by the citizen's no-entry strategy implies that conformist weights are added to the player's pay-offs. On the contrary, were the citizen willing to enter when the State adopts the mixed strategy (so that by entering she is equally compliant as when staying out), the State would become responsible for a sharp deviation from full compliance, for he could have chosen not to abuse at all. In that case, he would not have maximized the value of *T* as he possibly could have. This may not be the minimum value for *T*, but he has nonetheless produced a significant deviation from full compliance (proportional to the distance from the maximum value of *T* conditional on the citizen's choice). Thus, in this case the motivational weight of conformity cannot enter the utility functions of both players in all its strength.

The previous discussion illustrates a particularity in the way the State's conditional conformity index and reciprocally expected conformity index (as seen in the citizens' eyes) behave in games like the trust game, and in general in games where the strategy of one player would induce the same result whatever the behavior of the second player. The citizen's strategy ¬***e*** (the trustor's strategy in the trust game in general) in fact causes the same pair of payoffs whatever the reply of the State (the *trustee).* Hence the State by its behavior can't make any difference about the two pair of the players' payoffs that are possible when citizen-player chooses ¬***e*** , which both will be necessarily (1,1). Since T is a function of the material payoffs, also the value of T is thus invariant in the two states compatible with citizen's strategy ¬***e***. (Notice that in the sequential version of the trust game this is quite natural: by playing ¬***e*** the citizen, player A, stays out of the interaction and thus prevent the State form having any influence over the outcome of the game, which in fact is only one, whatever the decision of player B could have been.). This means that in our case the State, given the citizen's strategy ¬***e*** *cannot* do any better than to witness the first player bringing about the value 0 of function *T* representing the distribution principle of social welfare. Saying it differently, in case the citizen doesn't enter, no value higher than T = 0 does exist that can be obtained through a choice of the State. So, whatever the strategic choice deliberated by the State, it cannot induce any deviation from the maximum possible value of T, given ¬***e***. Neither of the

States' choices - let it be ***a*** or ¬***a*** - may deviate at any rate from the maximin possible value of T (= 0) given that the citizen's choice is ¬***e***. Thus for both the State's strategy choices, conformity will be as high as possible given the citizen's choice ¬***e***.

In terms of determinants of the State's conditional conformity index and expected reciprocity index (as seen in the citizen s' eyes) the differences between the T values determined by any State's strategy choice and the maximum possible value of T (conditional on the given the citizen's choice ¬***e***) are thus zero:

$$T(\boldsymbol{a}, \neg\boldsymbol{e}) - T^{MAX}(\neg\boldsymbol{e}) = 0, \qquad T(\neg\boldsymbol{a}, \neg\boldsymbol{e}) - T^{MAX}(\neg\boldsymbol{e}) = 0.$$

This is true for any pure or mixed strategy of the State (e.g. included any probabilistic combination of ***a*** and ¬***a***) granted that the citizen stays out.

This entails – and this is the peculiarity in how the indexes behave to be pointed out here – that the State's conditional deviation degree and the State's expected reciprocal deviation degree in the case under consideration are indefinite. In fact as far as no strategy of the State, given ¬***e***, may induce any difference with respect the value of T, this also entails that the Max and Min value of function T are even, given ¬***e*** (i.e. $T^{MAX}(\neg\boldsymbol{e}) = T^{MIN}(\neg\boldsymbol{e}) = 0$). So that their difference reported at the denominator is nil (i.e. neither the numerator nor denominator may report any distance from the maximum value of T given ¬***e***). Hence both the deviation degree and the reciprocally expected deviation degree are necessarily 0/0, namely indefinite. But of course this occurs because there is no proper sense in normalizing the measure of deviation from the max value of T given ¬***e*** with respect to the interval from 0 to -1, by taking it as a fraction of the distance between the maximum and the minimum value of T in cases where this distance is nil. In these case simply the fraction is meaningless.

Thus in this and all the analogous cases in which, given a certain adversary's choice, the maximum and minimum value of T determined by a player's choice scores difference equal to zero, we will assume that the degree of deviation from the maximum value of T due to this player's choices is simply represented by the *absolute value* of the difference between the T value determined by the player's choice (given the adversary's choice) and the maximum T value possible given that adversary's choice. Notice however that, because the T value is identical for all this players' choices given the adversary's behavior, the deviation is necessarily nil for that player and hence also this deviation measure - even without normalization – is necessarily 0. Thus the conformity indexes cannot be but 1.

Coming back to the trust game of Figure 3, by considering first the psychological utilities of the State, when the citizen is predicted to play ¬***e***, then the State would score full conformity both playing ***a*** or ¬***a***. But only when the State believes that the citizen predicts that he (the State) plays ***a*** then the citizen's reciprocal conformity would be full by using ¬***e***. In fact in case the State believed to be predicted to use ¬*a*, then the citizen's not entering choice would minimize T, and then the citizens reciprocally expected conformity would be 0. Thus under the strategy combination (¬***e***, ***a***), represented through first and second order beliefs of the State, the State's conformity index and the citizen's reciprocal expected conformity index equal 1. So that the weight λ fully enters the psychological payoff of the State. On the contrary under the combination (¬***e***, ¬***a***) - again seen through the State's beliefs - the citizen's reciprocal conformity index equals zero, what would nullify the weight λ in the State's psychological payoff.

As well, coming now to the citizen's psychological utilities, if the State is predicted to use the strategy ***a***, then the citizen's strategy ¬***e*** scores full conditional conformity, since by playing ***e*** the citizen would induce a lower T value and no other citizen strategy than ¬***e*** can induce a higher T value. Otherwise, if the citizen believes that the State predicts that she uses ¬*e*, then the State's reciprocal conformity expected in case the State is predicted to use ***a*** or ¬***a*** *is even (as high as possible in these contingency)*, i.e. the State's reciprocally expected conformity equals 1. Thus, given these citizen's conditional conformity and State's expected reciprocal conformity indexes for the combination (¬***e***, ***a)***, as seen through the citizen beliefs, the weight λ enters the psychological payoff of the citizen. This would not be the case if the citizen predicted that the State was to use ¬***a***. In fact, as far as the citizen plays ¬***e***, it is true that the States' expected reciprocal conformity index (as seen by the citizen) is even (and equal 1) for both the choices ***a*** and ¬***a***. But if she predicts ¬***a***, the citizen's conditional conformity of choosing ¬***e*** would be minimal (set to 0). So that the weight λ would be canceled in the citizen's psychological utility function for the (¬***e***, ¬***a)*** combination. Summing up, taking the game matrix line corresponding to the strategy ¬***e***, the weight λ enters the psychological payoffs of both the players only in the state represented by the bottom right cell.

What has been said till now is by no means conclusive about the existence of psychological equilibria based on conformist preferences in the one shot Trust Game. However it helps to understand how the psychological payoffs behave under different strategic and beliefs configurations. Psychological equilibria (in pure strategies) are then simply calculable. Inspection of Figure 4 shows that if the State is predicted to play strategy *a*, the citizen maximizes *T* by playing strategy ¬*e*. If this is known, the State also maximizes *T* by playing *a*, since neither strategy is

better or worse than *a* in order to maximize *T* from the State's point of view. Hence, in the bottom right cell of Figure 5 the psychological weight λ adds to each player's material pay-off. On the other hand, if the State is predicted to play ¬*a*, then the citizen maximizes *T* by choosing *e*. If this choice is also predicted by the State, his choice for maximizing *T* is ¬*a* as well. Consequently, in the top left cell of Figure 5 psychological weights λ are also present. If the State plays abuse (*a)*, the citizen will minimize *T* by entering (*e)*, which is also true if the same result is seen the other way round (given *e*, the State minimizes *T* by abusing with *a*). No weights must then be added in the top right cell of Figure 5. Lastly, if the State is predicted as not abusing, the citizen minimizes *T* by staying out with ¬*e*. Consequently, even though the State is maximizing *T* when he plays ¬*a* , a zero index of individual conformity (the citizen's) is sufficient to nullify the overall level of conformity. Moreover, when this is the case, no psychological conformity weights are implied in the players' pay-offs (see bottom left cell of Figure 5) .

Summing up, given the value λ = 2, we may see that, as far as only pure strategies are concerned, two Nash psychological equilibria do exists (***e, ¬a***) and (***¬e, a***). Thus even in the one shot game, the situation is ameliorated for not only the 'bad' equilibrium is now possible, but from the point of view of the solution determinateness the situation is also worsened as it isn't any unique. We don't bother here the reader with the existence of mixed-strategy-psychological-Nash equilibria in the one-shot Trust Game as they are mostly relevant to out argument in the context of the repeated Trust Game considered in the next section (where also many standard Nash equilibria are possible). It is within the perspective of the repeated Trust Game that we have to verify whether conformist preferences with an *ex ante* agreed principle of justice will simplify the equilibrium selection problem.

### b. <u>Conformist preferences in the iterated Trust Game: mixed strategies</u>

Now let us consider the repeated Trust Game. Recall that its pay-off space in terms of material utilities is the convex hull of all the linear (probability) combinations of the three pay-off vectors generated out of the pure strategy pairs of the basic Trust Game (see Figure 2). This is the same as representing the expected pay-offs of every possible pair of pure and mixed strategies of the two players in the basic Trust Game. In fact the player's *i* expected pay-off for a mixed strategy is formally the same as the *average pay-off* of the player's *i* repeated strategy that employs alternatively the two player's *i* pure strategies of the stage game with a given frequency, generating the three stage-game outcomes (1,1,), (4,4), (0,5) according to the frequency of the two players' choices. The cumulative pay-off of this repeated strategy, given a certain pure (or mixed) response

by the second player, can be equated to the average pay-off of a cycle along which player *i* gets each of the three stage-game payoffs a given number of times out of the total number of times defining the cycle (granted, of course, that during the game each repeated strategy pairs used by any player repeatedly enters a cycle with the same pattern of outcomes and the same average payoff value for the player that adopts it). It is thus simple to see that a State's mixed strategy that employs the two pure strategies $\neg\boldsymbol{a}$ and $\boldsymbol{a}$ with probability 0.25 and 0.75, respectively, against – to keep things simple – the citizen's pure entry strategy $\boldsymbol{e}$, affords the State and the citizen expected the pay-offs (0.25×4+0.75×5 = 4.75) and (0.25×4+0.75×0 = 1), respectively. This is equal to the average values attached to a repeated strategy whereby the State plays the stage-game strategy $\neg\boldsymbol{a}$ 75 per cent of the time and the stage-game strategy $\boldsymbol{a}$ 25 per cent of the time, assuming – to keep things simple again – that the citizen always responds with the stage-game strategy $\boldsymbol{e}$. It is obvious to see that in the one-shot Trust Game, no mixed strategy exists as a best response for the State. In the repeated Trust Game, however, one knows that this is no longer true. In fact, the State may create a reputation (along, for example, the first *N* repetitions of the game) to be a *type* that uses *the strategies* $\neg\boldsymbol{a}$ and $\boldsymbol{a}$ in a given frequency, such that the citizen's best response is 'always $\boldsymbol{e}$' until by repeated observations he realizes that the frequency is respected, but sanctioning by '$\neg\boldsymbol{e}$ forever' were it to become clear that the frequency is not respected. This induces the State to stick to its repeated strategy, mixing $\boldsymbol{a}$ and $\neg\boldsymbol{a}$ according to the given frequency.

One must, however, consider the pay-off space of the psychological game, which can be generated from that of the Trust Game when all of the expected pay-offs of mixed strategy pairs are accounted for. This repeated psychological Trust Game in pure and mixed strategies has the same material pay-off space as the repeated TG, wherein the average pay-offs of each repeated strategy – which employs the pure strategies of a player in a given frequency – is identical to the expected utility of the mixed strategy using the corresponding probability mixtures. Hence, one may ask what happens (under the psycological extension) to the mixed strategy equilibrium points of the corresponding standard repeated Trust Game.

Before answering that question, one must define a way to calculate the expected psychological utility of any mixed strategy. Let us take the point of view of the citizen (call him *A*) when she predicts that the State (call it *B*) will choose a mixed strategy, for example:

$$\sigma_B^{0.6} = \{(0.6, \neg a); (0.4, a)\}$$

*A* believes that if she enters by playing the pure strategy ***e***, two states (***e, ¬a***) and (***e, a***) may occur, so that two different values of the principle *T* – namely (9) and (-4) – can arise, each of them weighted with the probabilities 0.6 and 0.4 of the respective states. Hence, the expected Nash bargaining product generated by *B*'s mixed strategy $\sigma_B^{0.6}$, given *A*'s entrance, is 0.6×9 + 0.4×(-4) = 3.9, whereas if *A* does not enter, the expected *T* value is 0 as usual. Given $\sigma_B^{0.6}$, player *A*'s strategy ***e*** maximizes *T* in respect to any other pure or mixed strategy by *A*, whereas ***¬e*** minimizes it. It turns out that player *A*'s conformity indexes are 1 and 0 for her pure strategies, respectively.

On the other hand, player *B*'s conformity indexes are the following. Assuming that *B* believes *A* will enter, *B* does not maximize *T* by playing the strategy $\sigma_B^{0.6}$, because it is obvious that no-abuse would do better in terms of *T*. Nor does playing the mixed strategy minimize *T*, which in fact would happen by playing ***a***. As a result, *B*'s conformity index for strategy $\sigma_B^{0.6}$ is a somewhat intermediate value 0.61. But assuming that *B* believes that player *A* will not enter by ***¬e***. Then *B*'s mixed strategy $\sigma_B^{0.6}$ will maximize *T* no less than any other strategy by *B*. *B*'s conformity index under this hypothesis is thus 1. To conclude the example, consider *A*'s respective expected material pay-offs from playing ***e*** or ***¬e*** against the mixed strategy $\sigma_B^{0.6}$

$$EU_A(e, \sigma_B^{0.6}) = 2.4; \qquad EU_A(\neg e, \sigma_B^{0.6}) = 1$$

Similarly, player B's expected material pay-offs from playing the mixed strategy against the two pure strategies of player *A* are

$$EU_B(e, \sigma_B^{0.6}) = 4.4; \qquad EU_B(\neg e, \sigma_B^{0.6}) = 1$$

Since the conformity indexes of players *A* and *B* for the strategy pair $(e, \sigma_B^{0.6})$ are 1 and 0.61, respectively, the psychological conformity weight λ will enter the players' utility functions accordingly, that is, by a value (1)(0.61)λ. Given λ = 2 , the weight of the conformist motivation is 1.22, and the overall utility pay-offs of players *A* and *B* are 3.62 and 5.62, respectively.

In the repeated psychological Trust Game, these pay-offs correspond to the following pair of player *B* and player *A*'s repeated strategies: player *B* employs his pure strategies ***¬a*** and ***a*** repeatedly with frequency 0.6 and 0.4 respectively. By this repeated strategy, he tries to convince player *A* (or the sequence of short-run players who participate in the repeated game in the position of *A*) that he will stick to this frequency forever. Player *A* decides to play repeatedly her entry strategy ***e*** as long as she does not see player *B* employing *abuse* with a frequency higher than 0.4, but if this frequency is exceeded she will switch to '***¬e*** forever'. Since player *A*'s threat seems

convincing, player *B* plays *ad infinitum* his above-defined mixed repeated strategy. Assume that exactly 100 times are sufficient to say that the required frequency has been verified so that – if the players adopt the pair of repeated strategies described above – 100 times is a cycle that repeats more and more along the repeated game with always the same proportion of stage games with outcomes $(e, a)$ and stage games with outcome $(e, \neg a)$. The average pay-offs for this pair of repeated strategies – including the psychological component – is the vector (3.62, 5.62). It would seem to be a good incentive for player *A* to yield to player *B*'s mixed abuse strategy, but we will come back to this point a little later.

Following the method mentioned above, under the hypothesis $\lambda = 2$, it is in fact possible to account for the entire pay-off space of the psychological Trust Game, including mixed strategies as well: see Figure 6, where payoffs of pure and mixed strategies and their translations into the psychological game payoff space are represented. Up to the mixed strategy $\sigma_B^{0.39}$ no psychological utilities accrue to players and hence a region of the basic Trust Game payoff space does not translates into the psychological payoff space.

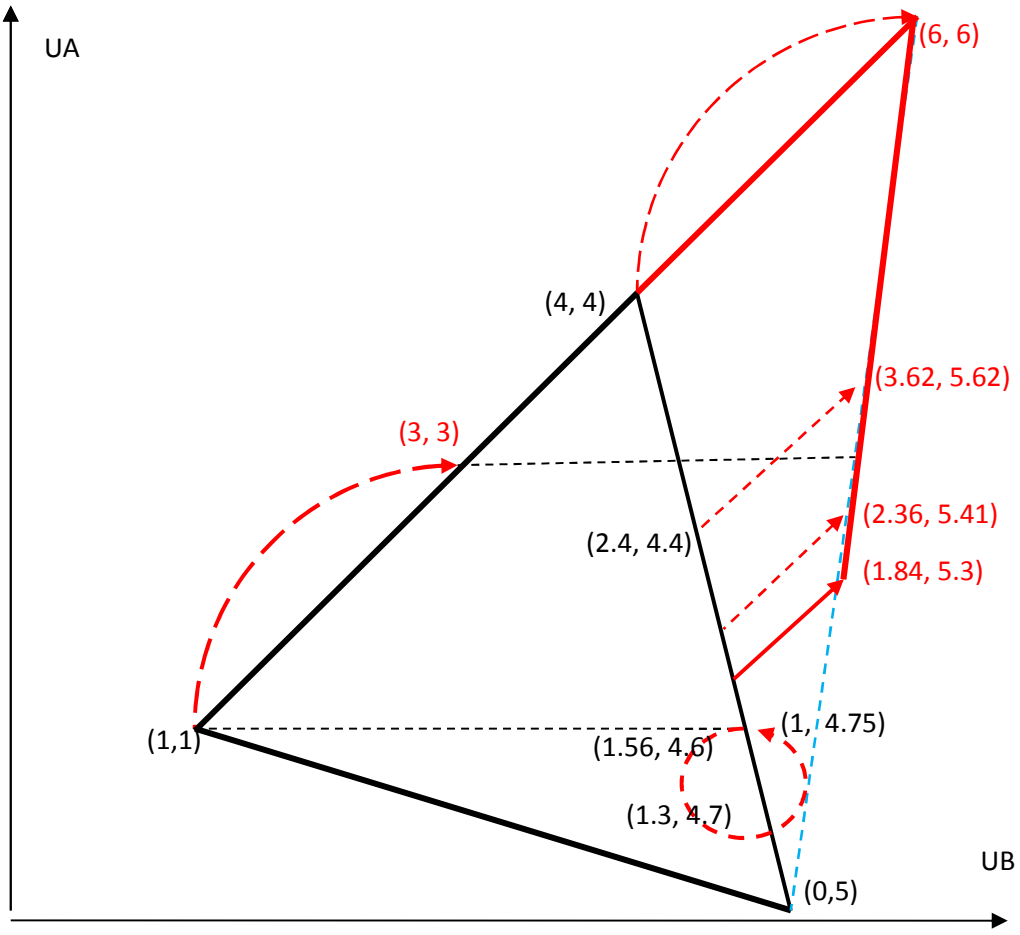


***Figure 6*** *The payoff space of the iterated psychological TG.*

First, let us note that the status quo point (1,1) – the only Nash equilibrium of the *basic one-shot* Trust Game and moreover an equilibrium of the *repeated* Trust Game – is translated in the northeast direction along the bisector to a point with overall utilities (3,3), which is also a psychological equilibrium of the new game. At the same time, thanks to the motivational conformist weights $\lambda = 2$, the outcome (4,4) where the Nash bargaining product is maximized translates in the northeast direction to the point (6,6), which is also a psychological equilibrium. Let us recall that both these psychological equilibria correspond to Nash equilibria of the repeated Trust Game, so that these two Nash equilibria are sure to be preserved under the pay-off change provided by conformist preferences.

In regard to player *B*'s mixed strategies, it can be seen that the entry strategy ***e*** of player *A* cannot be rewarded with any additional psychological conformist utility until the expected Nash Bargaining product – the expected value of *T* associated with any particular probability mixture of the two pure strategies $\neg\boldsymbol{a}$ and ***a*** – is no longer positive, granted player *A* uses ***e***. This necessarily happens until a mixed strategy associates the pure strategy $\neg\boldsymbol{a}$ with a probability high enough to give the respective *T* value (9) a weight able to counterbalance the *T* value of ***a*** (-4), so that the *T* expected value exceeds the *T* level fixed by the 'status quo' no-entry strategy (which is 0). Hence, within player *B*'s continuous set of probability mixtures of two pure strategies $\neg\boldsymbol{a}$ and ***a,*** the relevant threshold is fixed by player *B*'s mixed strategy that scores an expected Nash product no different from the *T* value of staying out. As long as this threshold is not exceeded, psychological pay-offs do not add any values to the material pay-offs of both players *A* and *B*, because entering by ***e*** minimizes the *T* value and exhibits zero conformity level. This is true also when player *B* adopts a mixed strategy that makes him partially, and hence positively, compliant. In fact until player *A*'s choice to enter by ***e*** exhibits a zero conformity index, the overall conformity level is also nil for both players and no psychological pay-offs can be added to their material pay-offs.

This does not mean that psychological utilities are not at work for these mixed strategies. Simply, the psychological component adds to the pay-offs of strategy pairs such as (*no entry*, *mixed strategy*), which is the same as for the strategy pair (*no entry*, *abuse*), namely (3,3). This means that the best responses for these cases is $\neg\boldsymbol{e}$, which gives player *A* an overall payoff 3, whereby player *B*'s mixed strategies and his pure strategy ***a*** become indifferent as they both give *B* the same overall payoff 3.

As an example, consider the mixed strategy $\sigma_B^{0.25} = \{(0.25, \neg\boldsymbol{a}); (0.75, \boldsymbol{a})\}$. The expected Nash bargaining product (the *T* value) is negative (−0.75) for the pair $(\boldsymbol{e}, \sigma_B^{0.25})$, whereas *T* is 0 if

player *A* chooses ¬***e***. It is thus obvious that *A* maximizes *T* by choosing ¬***e*** **,** with conformity index 1, whereas the conformity index for choosing ***e*** is 0. As a result, by entering with ***e***, player *A* can only get the expected overall pay-off 1, which – due to the probability mixture provided by $\sigma_B^{0.25}$ – is no different from the *material* pay-off of staying out. By staying out with ¬***e***, however, he gets an *overall* pay-off 3, because the psychological conformist weight 2 now adds to this strategy's material pay-off. Thus, *A*'s best response is obviously to stay out. As far as player *B* is concerned, the mixed strategy $\sigma_B^{0.25}$ against ***e*** gives a pay-off equal to its material pay-off 4.75**.** When player *A* does not enter against $\sigma_B^{0.25}$, *B*'s pay-off benefits from the psychological conformist component (becoming 3) as well as for any other choice (abusing or not abusing) by *B* when he knows that *A* will play no-entry.

Note the importance of the mixed strategy $\sigma_B^{0.25}$. This is player *B*'s Stackelberg mixed strategy that would correspond to the preferred (by the State) equilibrium strategy of the repeated Trust Game. It identifies exactly the equilibrium point of the repeated TG, that would be the most obvious choice from the point of view of player *B* were he able to select the solution of the game by himself. It is noticeable, however, that the pair (***e***, $\sigma_B^{0.25}$) is not an equilibrium in the psychological TG, even if player *B*'s material pay-off is high. Given the mixed strategy $\sigma_B^{0.25}$, neither is player *A*'s best response ***e***, nor is player *B*'s material pay-off 4.75 sufficient to make the strategy $\sigma_B^{0.25}$ preferred than ***a*** when *A* plays ***e,*** simply because, due to a sufficiently high λ associated with the psychological equilibrium in pure strategies (*entry*, *no-abuse*), playing ¬***a*** pays *B* more (namely 6).

The threshold that allows mixed strategies to gain support from psychological conformist utility is reached at the mixed strategy $\sigma_B^{0.307}$ ={(0.307, ¬***a***); (0.693, ***a***)}. Given this mixed strategy, the expected value of *T* is zero for any strategy choice by *A*, so that *A* is fully conformist by choosing either ***e*** or ¬***e*.** At the same time, playing the mixed strategy is partially conformist also for player *B*, because the minimization of the *T* value, given *A*'s entrance, would be obtained by playing ***a***. Hence, under the pair (***e***, $\sigma_B^{0.307}$), psychological utilities add to both the players' material pay-offs (1.3, 4.7) generating an overall pay-off vector (1.84, 5.31). It is important to note, however, that adding a bit of psychological utility does not mean that this strategy combination becomes a psychological equilibrium. Although it is true that player *B*'s mixed strategy $\sigma_B^{0.307}$ grants a positive overall pay-off to *A*'s entry strategy, player *A*'s overall pay-off from no-entry (i.e. 3) is still higher than the overall pay-off (1.84) from giving in to player *B*'s mixed strategy. This is due to the incomplete conformity level of strategy $\sigma_B^{0.307}$ when player *A* chooses ***e***. In fact *B*'s full conformity would be reached by the strategy ¬***a***, whereas $\sigma_B^{0.307}$ scores only the modest conformity index 0.31.

This affects the psychological conformist component of player *A*'s overall pay-off for strategy ***e***, which is lower than for ¬***e***.

Now let us consider mixed strategy $\sigma_B^{0.39}$ ={(0. 0.39, ¬***a***); (0. 61, ***a***)}. With this small increase in the probability of strategy ¬***a***, things finally seem to change. Player *A* with overall pay-off 2.36 benefits substantially from the psychological conformist utility of her entry strategy ***e***'. At the same time, as typically happens when a pure strategy is surpassed in its conformity index, player *A*'s conformity index of no-entry drops to zero, since choosing ¬***e*** given $\sigma_B^{0.39}$ would minimize the value of *T* in respect to the alternative entry strategy (and also any other mixed strategy). Hence, player *A*'s overall utility for the no-entry strategy ¬***e*** also dramatically drops to 1 (the material pay-off only). Moreover, for the pair (***e***, $\sigma_B^{0.39}$) , player *B*'s overall pay-off contains a substantial psychological conformist component such that his overall pay-off now reaches 5.41. If player *A* were to choose ¬***e***, however, player *B*'s pay-off would be reduced just to his material pay-off 1, since the conformity index of player *A*'s strategy ¬***e*** is zero (though *B*'s index remains positive). Note, nonetheless, that this does not imply that one has reached an equilibrium point. Even though entry is player *A*'s best reply to player *B*'s mixed strategy $\sigma_B^{0.39}$, this strategy is not reciprocally player *B*'s best response. The perfectly compliant strategy ¬***a*** would do better in terms of conformity index, scoring an overall pay-off 6 higher than the mixed strategy.

This suggests a general fact about the model. Let us consider again the mixed strategy $\sigma_B^{0.6}$ ={(0.6, ¬***a***); (0.4, ***a***)}. As we know, player *A*'s conformity index if she uses strategy ***e*** against $\sigma_B^{0.6}$ is 1, whereas the mixed strategy's conformity index is 0.61. The annexed overall pay-offs are (3.62, 5.62), respectively. Even though high psychological conformist utility enters both the players' pay-offs, this is not enough to define reciprocal best responses at (***e***, $\sigma_B^{0.6}$) since, given player *A*'s entry strategy, player *B*'s best reply is again no-abuse at all with its overall pay-off 6.

### c. <u>Equilibrium set of the psychological repeated Trust Game</u>

In order to give a general assessment of the two players' best reply sets in the psychological Trust Game, let us assume that λ is high enough for the pure strategy equilibrium (***e***,¬***a***) to exist. Let us call $E^{n|e}(\Pi_{A,B})$ the expected Nash Bargaining Product corresponding to player *B*'s n-ary mixed strategy $\sigma_B^n$ (where the index *n* corresponds to the probability weight assigned to the pure strategy ¬***a***) given player *A*'s strategy ***e.*** Hence, let $\Pi_{A,B}$ denote a generic Nash bargaining product.

Lastly, let's call 'status quo' the material pay-off granted by *A*'s pure strategy $\neg\boldsymbol{e}$. The relevant facts about the psychological Trust Game are the following.

- *Case 1*, $\forall\sigma_B^{\,n}$ with $n \geq 0$ s.t. $E^{n|e}(\Pi_{A,B}) < 0$, such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 > E(\Pi_{A,B})^n$, the pure strategy $\boldsymbol{e}$ does not add any psychological conformist utility to player *A*'s material pay-off, whereas the pure strategy $\neg\boldsymbol{e}$ adds the psychological conformity weight $\lambda$ to the 'status quo' material pay-off. Hence player *A*'s best reply is $\neg\boldsymbol{e}$ whereby *any* mixed strategy in this case is as good as strategy $\boldsymbol{a}$ to player *B*. The equilibrium for this case is the psychological equilibrium point ($\neg$**e**, $\boldsymbol{a}$). This equilibrium is weak since every mixed strategy in this case gives player *B* the same overall pay-off of $\boldsymbol{a}$.

- *Case 2*, $\forall\sigma_B^{\,n}$ with $0<n<1$ s.t. $E^{n|e}(\Pi_{A,B}) > 0$, such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 < E(\Pi_{A,B})^n$. Each pair ($\boldsymbol{e}$, $\sigma_B^{\,n}$) adds some psychological conformist utility to both players' material pay-offs, whereas the pure strategy $\neg\boldsymbol{e}$ reduces player *A* to the 'status quo' material pay-off. This follows from the minimal conformity index of strategy $\neg\boldsymbol{e}$, while in this case mixed strategies $\sigma_B^{\,n}$ have positive conformity indexes strictly less than 1. Thus for both players *A* and *B*, there is an intermediate overall index *F* of conditional and expected reciprocal conformity. In this case, player *A*'s best reply is strategy $\boldsymbol{e}$. Nevertheless, against strategy **e**, player *B*'s best is $\neg\boldsymbol{a}$. In other words, as little as player *B*'s psychological conformist utility of a mixed strategy $\sigma_B^{\,n}$ is positive, player *B*'s pure strategy $\neg\boldsymbol{a}$ against $\boldsymbol{e}$ (or whatever mixed strategy by player *A*) induces a psychological conformist pay-off higher than $\sigma_B^{\,n}$, so that player *B* has an incentive to deviate from $\sigma_B^{\,n}$ to $\neg\boldsymbol{a}$. When this occurs, player *A* obviously has no reason to change her choice, and the equilibrium point is (**e**, $\neg\boldsymbol{a}$).

- *Case 3*, for a single $0<n<1$ $\exists\,\sigma_B^{\,n}$ such that $E^{n|e}(\Pi_{A,B}) = 0$, such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 = E^{n|e}(\Pi_{A,B})$. In this case, both the strategy pairs ($\boldsymbol{e}$, $\sigma_B^{\,n}$) and ($\neg\boldsymbol{e}$, $\sigma_B^{\,n}$) add positive psychological conformist utility to the material pay-offs of both the players *A* and *B*. Nevertheless, player *A*'s overall pay-off gained from ($\neg\boldsymbol{e}$, $\sigma_B^{\,n}$) strictly dominates her overall pay-off gained from ($\boldsymbol{e}$, $\sigma_B^{\,n}$) since, whereas the two pure strategies $\boldsymbol{e}$ and $\neg\boldsymbol{e}$ score the same conformity index, the case of player *B*'s conformity indexes is different. Player *B* against $\neg\boldsymbol{e}$ cannot do any better than play $\sigma_B^{\,n}$ with conformity index 1, but given $\boldsymbol{e}$ the strategy $\sigma_B^{\,n}$ conformity index is strictly less than 1, which is the conformity index of his pure strategy $\neg\boldsymbol{a}$. Since the strictly less than 1 conformity index of strategy $\sigma_B^{\,n}$ directly depends on the required probability value *n*, which also affects the expected material utility of player *A* for ($\boldsymbol{e}$, $\sigma_B^{\,n}$), this correlation is crucial in this case. It

turns out that the greater player *A*'s pay-off gained from (***e***, ¬ ***a***) is, the smaller the probability required for the $\Pi_{A,B}$ indifference, but also the smaller the resulting player *B* conformity index for $\sigma_B^n$. Thus, player B's small conformity index at the same time affects negatively (via a small probability) player *A*'s material expected utility – since a small probability of (***e***, ¬ ***a***) will counterbalance its high pay-off – and also makes the strategy ***e*** psychological utility increasingly lower than the strictly dominant psychological utility of strategy ¬***e***. The resulting equilibrium point of this case is still (¬***e***, ***a***).

Boundaries between the three cases are established by the distribution of the material pay-offs associated with any mixed strategy, and in particular how much surplus it assigns to player *A*. As long as a mixed strategy overwhelmingly advantages player *B* in relation to player *A*, the *T* expected value of the mixed strategy pair (***e***, $\sigma_B^n$) cannot exceed that of player *A*'s staying out. This is not just because *A* is dissatisfied with his/her material outcome, but because of the insufficient conformity index of such mixed strategies. When a mixed strategy $\sigma_B^n$ instead offers a substantial share of the material surplus to player *A*, it becomes the most conformist solution, and then provides psychological utility to both the players against a loss of material pay-off to *B*. At this point, however, player *B* is able to compare the psychological utility of incomplete conformity against that of full conformity. It is evident that if the parameter $\lambda$ is high enough to guarantee the existence of the psychological equilibrium in pure strategies, then it is also true that player *B* will always prefer the pure strategy of full conformity.

This also depends, of course, largely on the $\lambda$ exogenous parameter of the two players (granted they are symmetric, which is not necessarily true). Were $\lambda$ too low, the situation would not change in regards to the basic and the repeated Trust Game. If, however, $\lambda$ is greater than player *B*'s pay-off difference between abusing and not abusing (given player *A*'s entry), its motivational effectiveness necessarily becomes maximal for the strategy of full conformity. In general, it biases the game towards excluding mixed strategies from giving rise to psychological equilibria. A look at the pay-off space reveals a single northeast vertex where both payers have highest pay-offs than anywhere on the eastern frontier where all the expected pay-offs generated by mixed strategies lie. In short, given its overall pay-offs, the pair (***e***, ¬***a***) strictly dominates any other strategy pair involving a mixed strategy $\sigma_B^n$ and player *A*'s entry strategy ***e.*** We have argued enough to state the following

*PROPOSITION I*

*Given a Trust Game with pure and mixed strategies, whereby a psychological game with conformist preferences is defined so that the motivational exogenous parameter $\lambda$ is great enough to guarantee the existence of a psychological equilibrium in correspondence to (**e**, ¬**a**), the game's psychological equilibria are only the two in pure strategy (**e**, ¬**a**) and (¬**e**, **a**), and no equilibrium points in mixed strategies exist. In particular, none of player B's mixed strategies is the best reply to player A's pure entry strategy **e**, even if the entry strategy **e** is player A's best reply tosome player B's mixed strategy*.

From this proposition comes the following

*COROLLARY*

In the repeated psychological Trust Game, psychological equilibria 'refine' the equilibrium set of the corresponding repeated Trust Game in a discontinuous way as a function of the increase in the motivational exogenous parameter $\lambda$.

- Given any $\lambda$ such that in the one-shot psychological Trust Game, there is no psychological equilibrium in correspondence with the pair (e, ¬a), the psychological equilibrium set of the repeated game is the same as the equilibrium set of the standard repeated Trust Game due to the sole effect of material pay-offs (see northeast boundary Z in Figure 7).

- If the value of $\lambda$ is such that in the one-shot psychological Trust Game player B's overall pay-off derived from the strategy combination (e, ¬a) is no different from the overall pay-off derived by B from the strategy combination (e, a) – so that a weak psychological equilibrium exists for (e, ¬a) – then in the corresponding psychological repeated Trust Game the psychological equilibria constituted by any mixed strategy $\sigma^{B}_{n}$ and the pure strategy e have all the same player B expected pay-offs, and thus they are all weak equilibria. Given the continuity of the probability mixture set over the two pure strategies ¬a and a, the value of $\lambda$ such that this is true is unique (see northeast boundary Y in Figure 7).

- If $\lambda$ is such that in the psychological one-shot Trust Game in correspondence to the pair (e, ¬a) there is a strong psychological equilibrium, then in the repeated psychological Trust Game there are no psychological equilibria in mixed strategies and the psychological equilibrium set dramatically shrinks to the only two pure strategy equilibrium points (e, ¬a) and (¬e, a). (See northeast boundary X of Figure 7).

The corollary is important, because it is in this context that we see our result. As far as the pay-off space of a one-shot basic Trust Game is concerned, mixed strategies are not equilibria. If B adopts a mixed strategy that induces A to enter, B immediately has an incentive to deviate to the abuse strategy since the mixed strategy is not the best reply to A's choice to enter. On the contrary, if the pay-off space is seen (as in the corollary) as the convex set of all the average pay-offs for repeated strategies in a repeated TG, then represented within this space may be the average pay-offs of player B's repeated strategies mixing the two pure strategies a and ¬a according to some pre-established frequencies.

Thus, if player *B* is able to accumulate a reputation of being a player that unfailingly plays one such strategy, he will have no reason to deviate if player *A* adopts a conditioned strategy of entrance like 'as long as my observations are compatible with the hypothesis that *B* is playing ***a*** and ***¬a*** according to the given pre-established frequency, I will continue to enter by **e,** but if I find that my observations are incompatible with that frequency, I will switch to **¬e** forever'. In fact, given player *A*'s conditioned entrance strategy, player *B* verifies that maintaining his reputation of being the type of player who uses the repeated strategy 'abuse no more than *x* per cent of the time, and no abuse for the rest of the time' is profitable since it allows him to gain a certain portion of the surplus. Summing up, player *B* has the incentive to keep abuses at a certain frequency in order to support his reputation of being the relevant type.

The situation changes significantly when the repeated psychological Trust Game is considered, however. In this case, a pay-off space identical to the convex hull of all the pay-off pairs deriving from pure strategy combinations in the one-shot psychological Trust Game is not completely generated by taking the set of all the *average* pay-off pairs given by combinations of the two players' (pure and mixed) repeated strategies (in fact payoffs spaces of Figures 6 and 7 have a non convex region along the dotted line from the payoff pair (0,5) to the payoff pair (1.84, 5.3)). What happens is that if player *B* has chosen a repeated mixed strategy whereby he has been able to accumulate a positive reputation that induces player *A* to enter for the first time, then he immediately recognizes the incentive to switch to a strategy that employs ***¬a*** with higher frequency.

This feature of the repeated psychological Trust Game completely changes the best response structure with regard to the standard repeated TG. In the standard case, player *B* has a clear incentive to maintain his strategy once he has been able to build up a reputation for being a mixed

*type*, since abusing less would give away a larger part of the surplus to player *A*, while abusing more would induce player *A* to carry out her sanction.

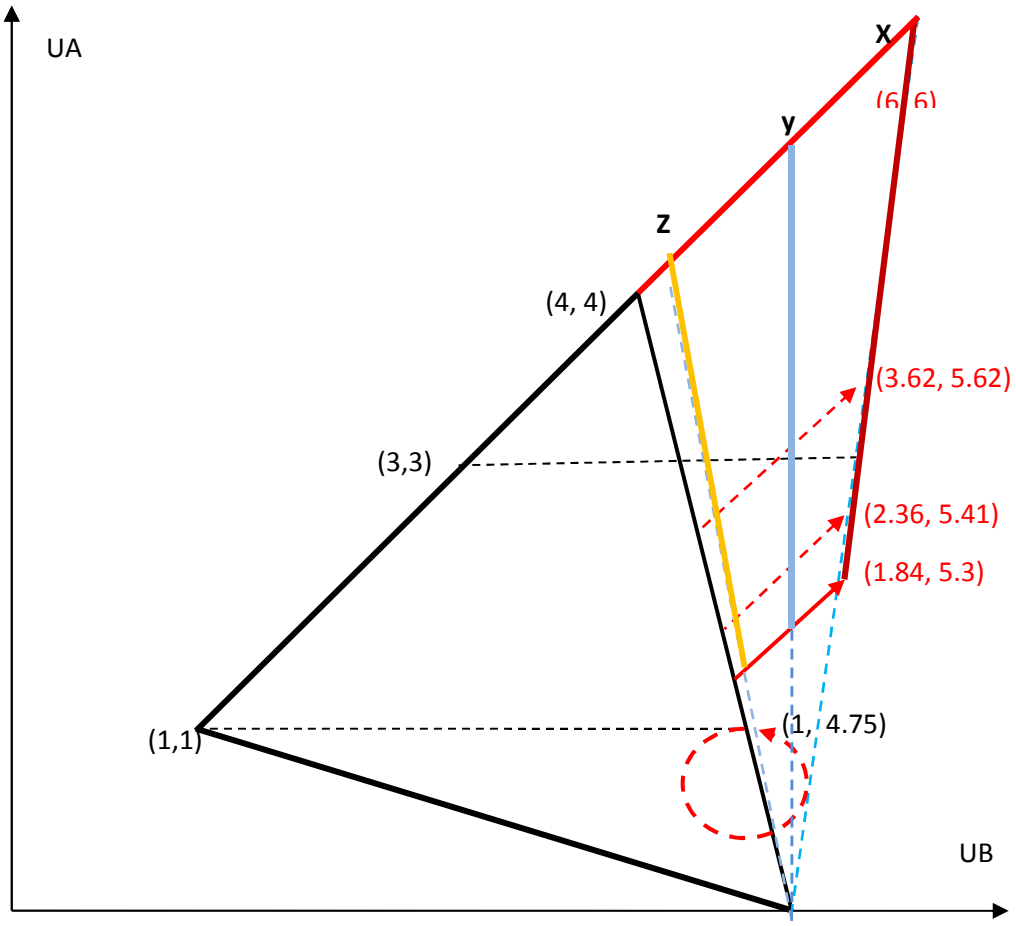


***Figure 7*** *Payoff spaces of the repeated psychological Trust Game under three values of the parameter* $\lambda$
*$\lambda$ <1 implies the NE frontier Z*
*$\lambda$ = 1 implies the NE frontier Y*
*$\lambda$ =2 implies the NE frontier X*

At the same time, player *A* has a strong incentive to monitor and sanction the relevant possible deviation by player *B*. In the repeated psychological TG, by contrast, player *B*'s best reply to player *A*'s entry is to deviate from any mixed strategy $\sigma_B^n$ to $\neg\boldsymbol{a}$. If, however, player *B* deviates to a strategy more concessive to him, *A* does not have any reason to punish him. Thus, the repeated mixed strategy equilibrium of the basic repeated Trust Game is destabilized. Summing up, any mixed strategy by player *B* that induces player *A* to enter, according to player *B*'s point of view is dominated by the pure strategy 'always $\neg\boldsymbol{a}$', so that a rational player *B* would never strive after a reputation such as being committed to the mixed strategy $\sigma_B^n$. From the outset, he would prefer to develop the dominant reputation of being an 'always $\neg\boldsymbol{a}$' player**.**

From this, the conclusion follows that even though generating a psychological game from a basic Trust Game enables us to determine new equilibrium points (in other words, to pass from only one equilibrium to at least two), when the change involves a step from the one-shot Trust Game to the repeated TG, transforming the pay-off space by means of conformist preferences has a powerful effect in reducing the psychological equilibria to a subset of the Nash equilibria.

## 5. Conclusion

In this paper we presented an application to the relation State/citizen of the theory of conformist preference, which is a behavioral game theoretical reformulation of John Rawls' theory of the sense of justice as a mechanism stabilizing just institutions once the original position has been exited – an explanation which is fully consistent with the theory of *ex ante* choice.

We started from the acknowledgment that in general, in a context of constitutional choice where agents confront one another in a state of nature, it can be shown that an egalitarian constitution is agreed (Binmore, 2005). Such egalitarian constitution is not only the outcome of cooperative bargaining, it is also the result of an equilibrium selection device, which ends up in a (self-sustaining) Nash equilibrium. This solution, however, is a necessary but not sufficient condition for the *ex post* compliance and stability of the social contract, as we have highlighted in the discussion about Binmore's result in the first section.

The introduction of the sense of justice and of conformist preferences brought us to an *ex post* model where the social contract, understood as a rational device with which to derive social norms and institutions through an original decision respecting basic ethical conditions (impersonality, impartiality and empathy), does not only bring to the *ex ante* selection of a fair equilibrium, but it also creates endogenously the motivations for its compliance and stability.

This is a great improvement on the previous social contract theories that, because they assumed that agreement on the social contract is a cooperative choice, were unable to solve the problem of its implementation in a non-cooperative context.

The main result of this paper is that an egalitarian social contract matters because it shapes preferences. If players have agreed rationally on a certain constitution, this affects their preferences, admitted that they expect reciprocity from others in complying with the same institution. Under expectations of reciprocity, preferences incorporate a desire to conform with agreed principles, which is a direct function of the extent to which an agent is conforming given his expectation of

others' behavior, and the extent to which he expects reciprocal conformity by others given their belief in his own action. Impartial agreement on a principle lays the bases; then mutual expectations are essential in shaping a preference for conformity. Hence the social contract eminently affects our preference for compliance; and the role for normativity of the social contract is adequately rescued.

This is not the whole story, however. The force of this preference can be great or small according to our attitude and disposition towards compliance with agreed rules. This can be seen as a parameter exogenous with respect to the social contract on specific institutions. It is contingent on cultural and social evolution, so that in a given context and phase the moral force of the desire to conform with just principles may not be descriptively strong enough to override other incentives (for example mere self-interest). Social and cultural evolution may be decisive in causally affecting the motivational force of a normative argument, without denying that the impartial agreement is a source of prescriptive and universalisable reasons to act.

Therefore, there are important consequences of the results set out in this paper for the fields of Constitutional Political Law and Economics. Going back to Buchanan (1975), we find this unresolved question: "Democracy may become its own Leviathan unless constitutional limits are imposed and enforced. (…) Can modern man, in Western democratic society, invent or capture sufficient control over his own destiny so as to impose constraints on his own government, constraints that will prevent the transformation into the genuine Hobbesian sovereign?"[7].

Our answer is that we need a constitution based on a Rawlsian social contract, which does not only display the property of fairness, but also of self-enforceability, because it is able to generate endogenously the motivations for its stability, while other institutions that do not rely on the sense of justice need a high degree of external enforcement since they do not have impact on preferences and motivations.

This helps explaining many historical examples of citizens punishing the governors' abuse of authority, even when their material payoff cannot explain their behavior.

When a society is based on such a fair Rawlsian constitution, chosen behind the veil of ignorance, it is capable of preventing the diversion of power, while those societies that do not have this constitutional basis cannot enable security on the part of citizens and therefore can easily deviate toward the abuse of authority.

[7] Buchanan (1975), 7.9.42.

## Appendix – Conformist Preference Model

Players have *two* kinds of preferences defined over states of affairs resulting from their interaction, which are both capable of motivating their actions. On one hand (more basic), the first kind of preferences is based on the description of states of affairs $\sigma$ brought about by their interaction *as consequences*, and their preferences regarding consequences are called *consequentialist*. These may be not only typical self-interested preferences but also altruistic ones. This part of the argument is by no means new. The new part instead concerns *conformist preferences*. Players also have preferences defined over states of the affairs $\sigma$ resulting from their interaction but described as just *combination of actions.* When these states of affairs are qualified in terms of their consistency with an *ex ante* agreed ethical norm preference over them are *conformist* – where "consistency" is defined as how far the players' strategy choices (jointly a state) are from the set of actions that would completely fulfil the agreed ethical norm of equity. By norm we mean a principle of justice for the distribution of material utilities coinciding with the *ex ante* social contract.

Let us assume that players have agreed on a social contract concerning the principle of justice that should govern cooperation in society. Conformist preferences may now enter the picture. Intuitively speaking, a citizen will gain intrinsic utility from simply complying with the principle, if the same citizen expects that in doing so she will be able to contribute to fulfilling the distributive principle, and taking into account that she expects the other citizens (or the State) also to contribute to fulfilling the same principle, given their expectations.

A complete measure of the player preferences is an overall utility function combining material utility, derived from her consequentialist preferences, with the representation of her conformist preferences represented by the conformist-psychological component of her utility function. The overall utility function of player $i$ with reference to the state $\sigma$ (understood as a strategy combination of player $i$ strategy $\sigma_i$ and the other players' strategies $\sigma_{-i}$), is the following

$$V_i(\sigma) = U_i(\sigma) + \lambda_i F[T(\sigma)] \qquad 1)$$

where:

i. $U_i$ is player $i$'s material utility for the state $\sigma$
ii. $\lambda_i$ is an exogenous parameter $\lambda_i \leq 0$;
iii. $T$ is a fairness principle defined for the state $\sigma$;
iv. $F$ is a compounded index expressing the agent $i$'s conditional conformity and her expectation of reciprocal by any other player $j$ with respect to the principle $T$ for each state $\sigma$.

Let's concentrate on the conformist part of the utility function. *First* (as it can be seen within the most internal brackets), there is a norm $T$, a social welfare function that establishes a distributive principle of material utilities. Players adopt $T$ by agreement in a pre-play phase and employ it in the generation of a consistency ordering over the set of possible states $\sigma$, each seen as a combination of individual strategies. The highest value of $T$ is reached in a situation $\sigma$ where material utilities are distributed in such a way that they are mostly consistent with the distributive principle $T$ within the available set of alternatives. Note that what matters to $T$ is not "who gets how much" material pay-off (the principle $T$ is neutral with respect to individual positions), but how utilities are distributed across players. Satisfaction of the distributional property is the basis for conformist preferences. As we are looking for a contractarian principle of welfare distribution, let us assume that $T$ coincides with the Nash bargaining function taking the stay out outcome of the trust game as the *status quo.*

*Agreed principle of fair welfare distribution $T$:*

$$T(\sigma) = N(U_1, \dots, U_n) = \prod_{i=1}^{n}(U_i - d_i) \quad 2)$$

*Second*, a measure of the extent to which, given the other agents' expected actions, the first player by her strategy choice contributes to a fully fair distribution of material pay-offs in terms of the principle $T$. This may also be put in terms of the extent to which the first player is *responsible* for a fair distribution, given what (she expects that) the other player will do. It is a *conditional conformity index* assuming values from 0 (no conformity at all, when the first player chooses a strategy that minimizes the value of $T$ given his/her expectation about the other strategy choice) to 1 (full conformity, when the first player chooses a strategy that maximizes the value of $T$ given the other player's expected strategy choice) with the following form.

*Player $i$'s conditional conformity index:*

$$1 + f_i(\sigma_{ik}, b_i^1) \quad 3)$$

This index takes its values as a function of $\boldsymbol{f_i}$ which in turn varies from 0 to -1 and measures player $i$'s *deviation degree* from the ideal principle $T$ by making her choice conditional on her expectation about player $j$'s behavior

*Player $i$'s deviation degree:*

$$f_i(\sigma_{ik}, b_i^1) = \frac{T(\sigma_{ik}, b_i^1) - T^{MAX}(b_i^1)}{T^{MAX}(b_i^1) - T^{MIN}(b_i^1)} \quad 4)$$

where $b_i^1$ is player $i$'s belief concerning player $j$'s action, $T^{MAX}(b_i^1)$ is the maximum value of the function $T$ due to whatever feasible strategy player $i$ may choose given her belief about player $j$'s choice, $T^{MIN}(b_i^1)$ is the minimum value of the function $T$ due to whatever feasible strategy player $i$ may choose given her belief about player's $j$ choice, and $T(\sigma_{ik}, b_i^1)$ is the actual value of $T$ due to player $i$ adoption of her k-ary strategy $\sigma_{ik}$ given her belief about player $j$'s choice.

*Third*, a measure of the extent to which the *other* player is expected to contribute to a fair payoff distribution in terms of the principle $T$, given what he is expected to expect from the first player's behaviour. This may also be put in terms of the (expected) *responsibility* of the *other* player for generating a fair allocation of the surplus, given what he (is believed to) believes. This measure consists of a *reciprocally expected conformity index* assuming values from 0 (no conformity at all, when the *other* player is expected to choose a strategy that minimizes $T$ given what he expects from the first player) to 1 (full conformity, when the *other* player is expected to maximize the value of $T$ given what he expects from the first players). It is formally very similar to the conditional conformity index of the first player, i.e.

*Player $j$'s reciprocal expected conformity index:*

$$1 + \tilde{f}_j(b_i^2, b_i^1) \quad 5)$$

In fact it is as well a function of $\tilde{f}_j$, the *expected player $j$'s degree of deviation* from the ideal principle $T$, which also varies from 0 to –1 as is also normalized by the magnitude of the difference between player $j$'s full conformity and no conformity at all, given what he believes (and player $i$ believes that he believes) about player $i$'s choice, i.e.

*Expected player $j$'s deviation degree:*

$$\tilde{f}_i(b_i^1, b_i^2) = \frac{T(b_i^1, b_i^2) - T^{MAX}(b_i^2)}{T^{MAX}(b_i^2) - T^{MIN}(b_i^2)} \quad 6)$$

where $b_i^1$ is player $i$'s *first order* belief about player $j$'s action (i.e. formally identical to a strategy of player $j$), $b_i^2$ is player $i$'s *second order* belief about what player $j$ believes about the action adopted by player $i$, while $T^{MAX}(b_i^2)$ and $T^{MIN}(b_i^2)$ are defined as above but in relation to player $i$'s second order belief.

*Fourth*, there is an exogenous parameter $\lambda$ ($\lambda \geq 0$) representing the motivational force of the agent's psychological disposition to act on the motive of reciprocal conformity with an agreed norm. This is a psychological parameter representing how strong the *sense of justice* or the "desire to be just" has grown up for an individual in a given population; it may be taken as dependent on exogenous variables like as the development of the affective capacity to act upon one's principles and duties that comes from lower level domain of interaction (as in Rawls' theory of moral development, the family and the circle of friends and small scale associations). Notice however that in the model it doesn't operates as such but as only once the agreement over $T$ is given and as it is weighted by the measure of reciprocal conformity.

In fact steps *two* and *three* coalesce in defining an overall index $F$ of conditional and expected reciprocal conformity for each player in each state of the game. This index operates as a *weight* on the parameter$\lambda$, deciding whether it will actually affect or not (and, if so, to what extent) the player's pay-offs. Thus the complete psychological component of the utility function representing conformist preferences is

$$\lambda_i\left[1 + \tilde{f}_j(b_i^2, b_i^1)\right][1 + f_i(\sigma_{ik}, b_i^1)] \qquad 7)$$

which reduces to the following cases:

$$\lambda[(1-x)(1-y)] = \lambda$$

i) $\lambda[(1-x)(1-y)] = \lambda$, since both *x* and *y* are 0, if player $i$ doesn't deviate and expects that player $j$ doesn't deviate at all from complete conformity;

ii) $\lambda[(1-x)(1-y)] = \alpha\lambda < \lambda$, where $\alpha < 1$ since $0 < x < -1$ and/or $0 < y < -1$, if player $i$ partially deviates and/or expects player $j$ to partially deviate from complete conformity;

iii) $\lambda[(1-x)(1-y)] = \alpha\lambda = 0$, since at least one (or both) of $x$ or $y$ are $-1$, if player $i$ does not conform at all and/or expects that player $j$ doesn't conform at all.

## References


Binmore, K. (2005) *Natural Justice*, Oxford: Oxford University Press.

Buchanan, J. (1975), *The Limits of Liberty: Between Anarchy and Leviathan*, Library of Economics and Liberty [Online] available from http://www.econlib.org/library/Buchanan/buchCv7.html.

Fudenberg, D. (1991) 'Explaining Cooperation and Commitment in Repeated Games', in J.J. Laffont (ed.), *Advances in Economic Theory, 6th World Congress*, Cambridge: Cambridge University Press.

Fudenberg, D. and D. Levine (1989) 'Reputation and Equilibrium Selection in Games with a Patient Player', *Econometrica*, **57**, pp. 759–778.

Gauthier, D. (1986), *Morals by Agreement*, Oxford: Clarendon Press.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989) 'Psychological Games and Sequential for Non-Cooperative Games', *International Journal of Game Theory*, **5** (1975), pp. 61–94.

Grimalda, G. and L. Sacconi (2005) 'The Constitution of the Not-for-Profit Organisation: Reciprocal Conformity to Morality', *Constitutional Political Economy*, **16** (3), pp. 249–276.

Rabin, M. (1993) 'Incorporating Fairness into Game Theory', *American Economic Review*, **83** (5), pp. 1281–1302.

Rawls, J. (1971) *A Theory of Justice*, Oxford: Oxford University Press.

Sacconi L. and M. Faillo (2010) 'Conformity, Reciprocity and the Sense of Justice. How Social Contract-based Preferences and Beliefs Explain Norm Compliance: the Experimental Evidence', *Constitutional Political Economy*, **21** (2), pp. 171–201.

Sacconi, L., Faillo, M. and Ottone S. (2011), 'Contractarian Compliance and the "Sense of Justice": A Behavioral Conformity Model and Its Experimental Support', *Analyse & Kritik*, 01/2011, pp. 273-310.